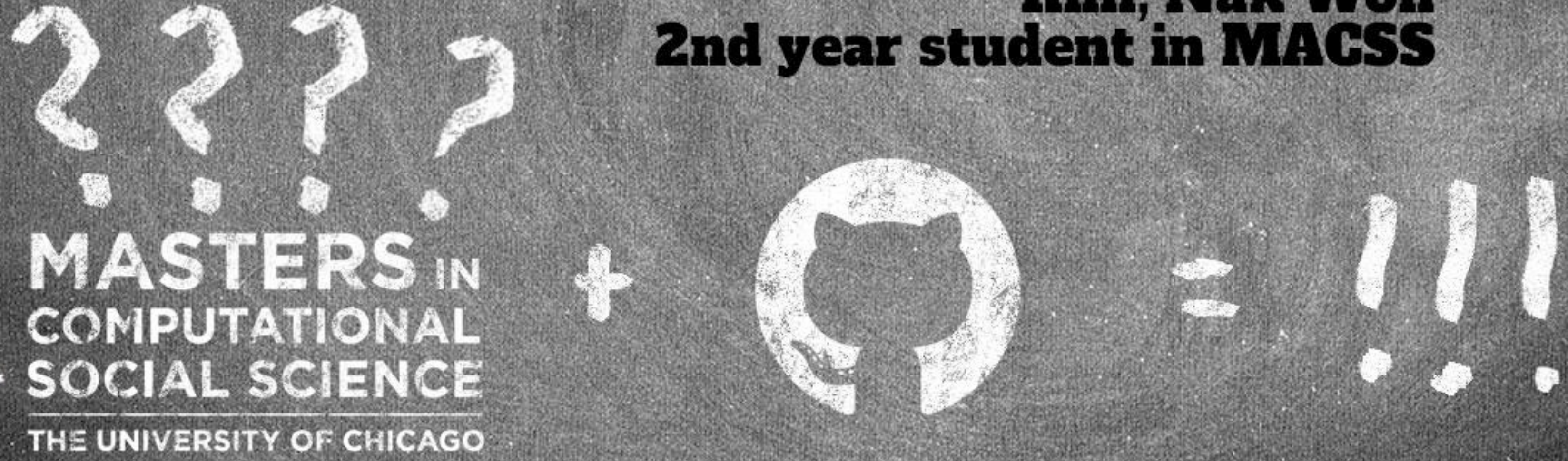
CSS Workshop on CSS Workshop

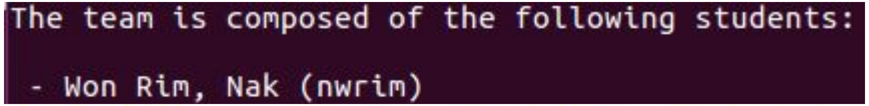
Rim, Nak Won
2nd year student in MACSS


????
MASTERS IN COMPUTATIONAL SOCIAL SCIENCE
THE UNIVERSITY OF CHICAGO
+
=
!!!

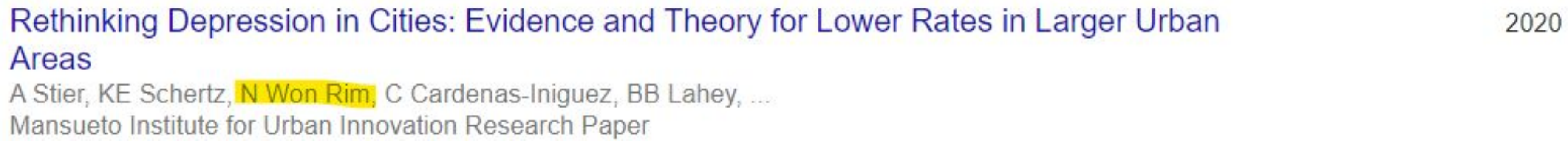
Who am I?

I am:

1. A MACSS 2nd year student
2. working on projects in the Knowledge Lab (Prof. Evans) and the Environmental Neuroscience Lab (Prof. Berman)
3. who got extremely bored in a 14-hour flight
4. and has a name that confuses many systems (or maybe I am just Korean)



```
The team is composed of the following students:

 - Won Rim, Nak (nwrim)
```



Rethinking Depression in Cities: Evidence and Theory for Lower Rates in Larger Urban Areas 2020

A Stier, KE Schertz, N Won Rim, C Cardenas-Iniguez, BB Lahey, ...

Mansueto Institute for Urban Innovation Research Paper

Goal of this little presentation

> The book fascinated him, or more exactly it reassured him. In a sense **it told him nothing that was new, but that was part of the attraction.**
>
> \- George Orwell, *1984*, Chapter 2, IX

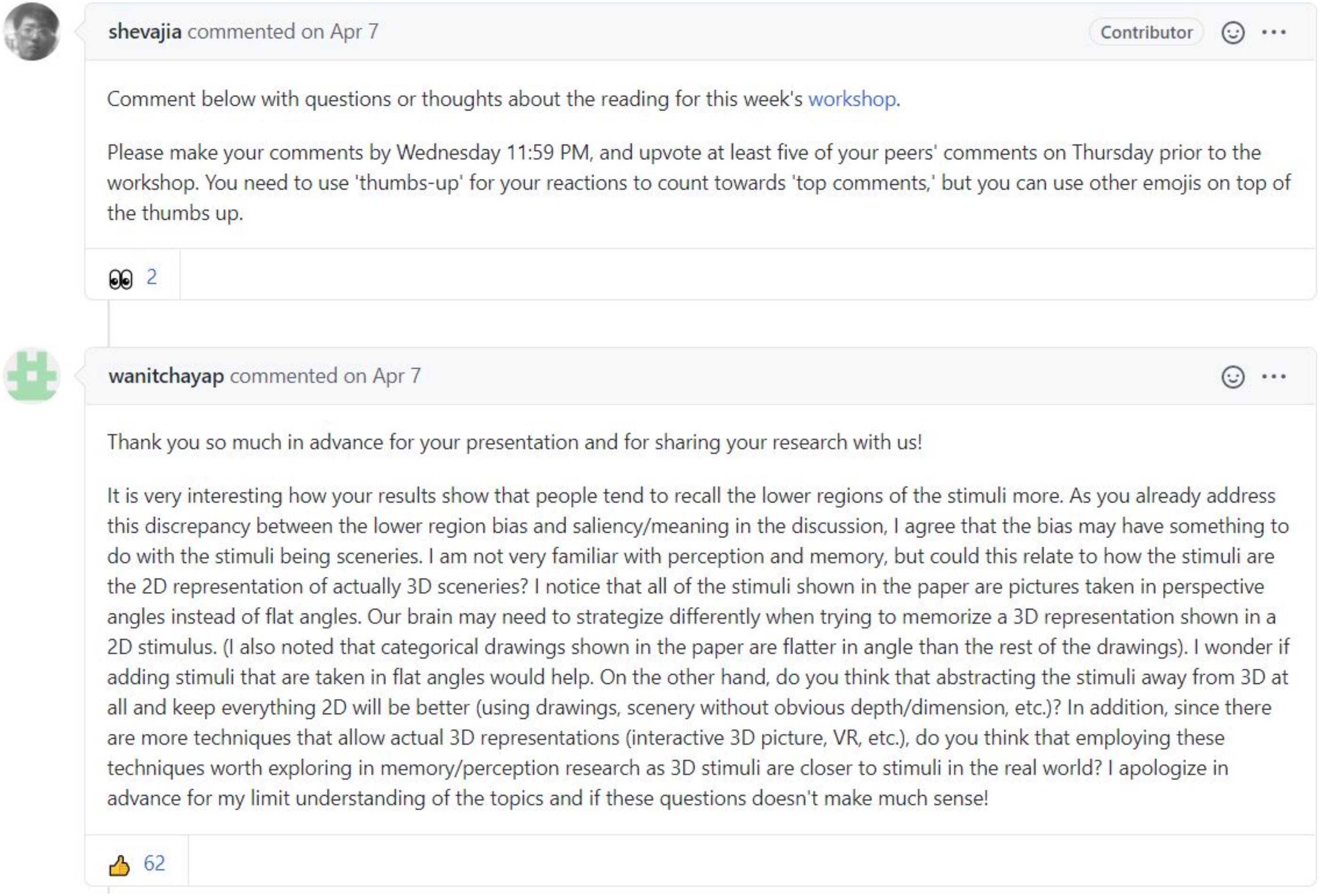
Table of Contents

- Some Background Information
- Workshop Questions Similarities
- Statistical Analysis of the “Primacy Effect”

Some Background Information

Our Ritual for the CSS Workshop

1. A preceptor sends relevant materials (usually a scholarly article or two) to students prior to the workshop each week
2. **Students read the material and post questions on a Github issue thread** (Deadline: Wednesday midnight)
3. **Students read each other's questions and upvote (‘thumbs-up’) questions that they think are good**
4. **Most upvoted questions are asked in the workshop**



shevajia commented on Apr 7

Comment below with questions or thoughts about the reading for this week's workshop.

Please make your comments by Wednesday 11:59 PM, and upvote at least five of your peers' comments on Thursday prior to the workshop. You need to use 'thumbs-up' for your reactions to count towards 'top comments,' but you can use other emojis on top of the thumbs up.

👀 2

wanitchayap commented on Apr 7

Thank you so much in advance for your presentation and for sharing your research with us!

It is very interesting how your results show that people tend to recall the lower regions of the stimuli more. As you already address this discrepancy between the lower region bias and saliency/meaning in the discussion, I agree that the bias may have something to do with the stimuli being sceneries. I am not very familiar with perception and memory, but could this relate to how the stimuli are the 2D representation of actually 3D sceneries? I notice that all of the stimuli shown in the paper are pictures taken in perspective angles instead of flat angles. Our brain may need to strategize differently when trying to memorize a 3D representation shown in a 2D stimulus. (I also noted that categorical drawings shown in the paper are flatter in angle than the rest of the drawings). I wonder if adding stimuli that are taken in flat angles would help. On the other hand, do you think that abstracting the stimuli away from 3D at all and keep everything 2D will be better (using drawings, scenery without obvious depth/dimension, etc.)? In addition, since there are more techniques that allow actual 3D representations (interactive 3D picture, VR, etc.), do you think that employing these techniques worth exploring in memory/perception research as 3D stimuli are closer to stimuli in the real world? I apologize in advance for my limit understanding of the topics and if these questions doesn't make much sense!

👍 62

It is publicly online and digitized...

... so we can analyze it

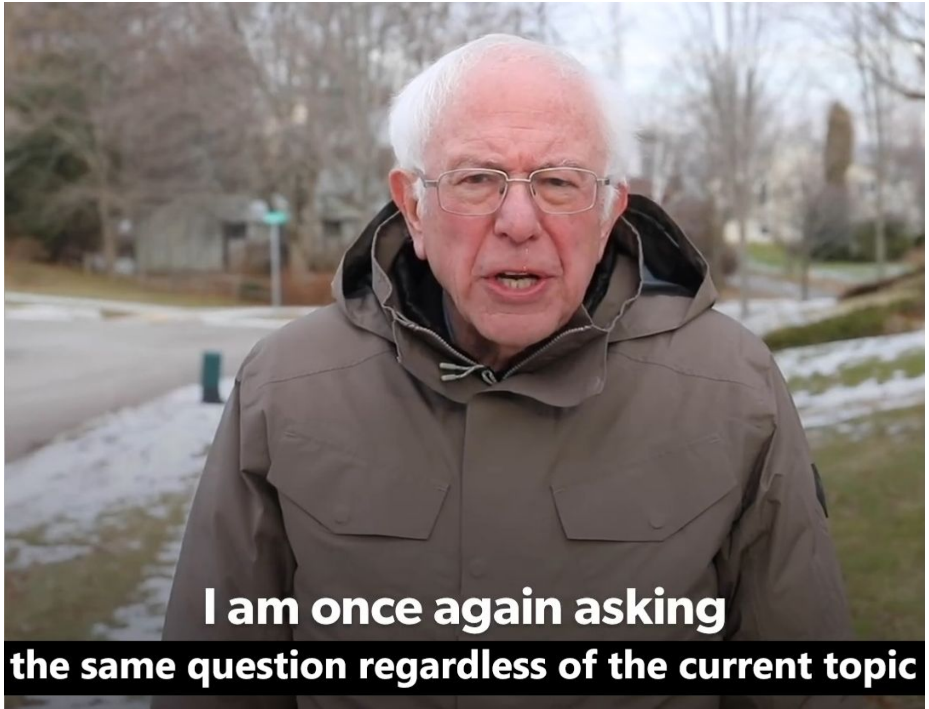
- We had **19 workshops** in 2019-2020 and **1138 comments** (excluding ones posted after the deadline; 304 comments)
- This forms a small corpus! Not big enough to do word embedding or train NN robustly (even topic modeling is a bit hard), but still.

Workshop Questions Similarities

Workshop - Best Case Scenario

Students read the text carefully and share their deep thoughts and questions, **producing very distinct questions for workshops with different topics**

Workshop - Worst Case Scenario



Method

1. Aggregated all comments for each workshop (i.e., 19 documents)
2. Calculated the **TF-IDF vectors** (with lemmatization) of the documents
3. Calculated the **cosine similarities** between each workshop vectors (simple, but effective way of computing document similarities - used in researches e.g., Thompson & Henry, 2018)
4. Ran **agglomerative (hierarchical) clustering** on the resulting similarities

Thompson, N., & Hanley, D. (2018), Science Is Shaped by Wikipedia: Evidence From a Randomized Control Trial . *MIT Sloan Research Paper, 5238-17*. http://dx.doi.org/10.2139/ssrn.3039505

Bag of Word Vectors (TF-IDF is weighted BOW)

- Example Documents:
  - 'But man is not made for defeat'
  - 'A man can be destroyed, but not defeated'
  - 'So we beat on, boats against the current'

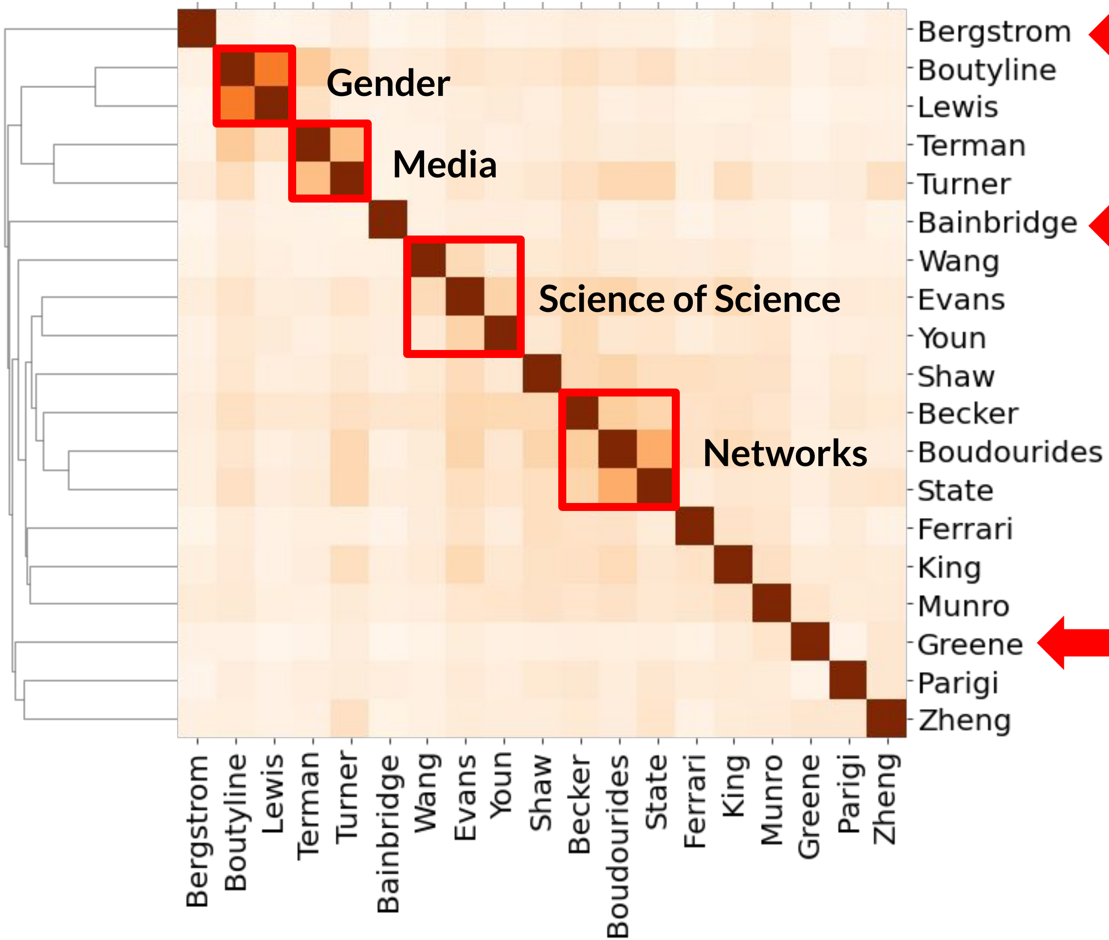
man	defeat	destroy	beat	boat	current
1	1	0	0	0	0
1	1	1	0	0	0
0	0	0	1	1	1

Cosine Similarity

$$\frac{A \cdot B}{\|A\| \times \|B\|}$$

0.816

0.000



Most Distinct ("Calling Bullshit")

Second Most Distinct ("Drawings of Real-world Scenes During Free Recall Reveal Detailed Object and Spatial Information in Memory")

Third Most Distinct ("Big Data in Medicine")

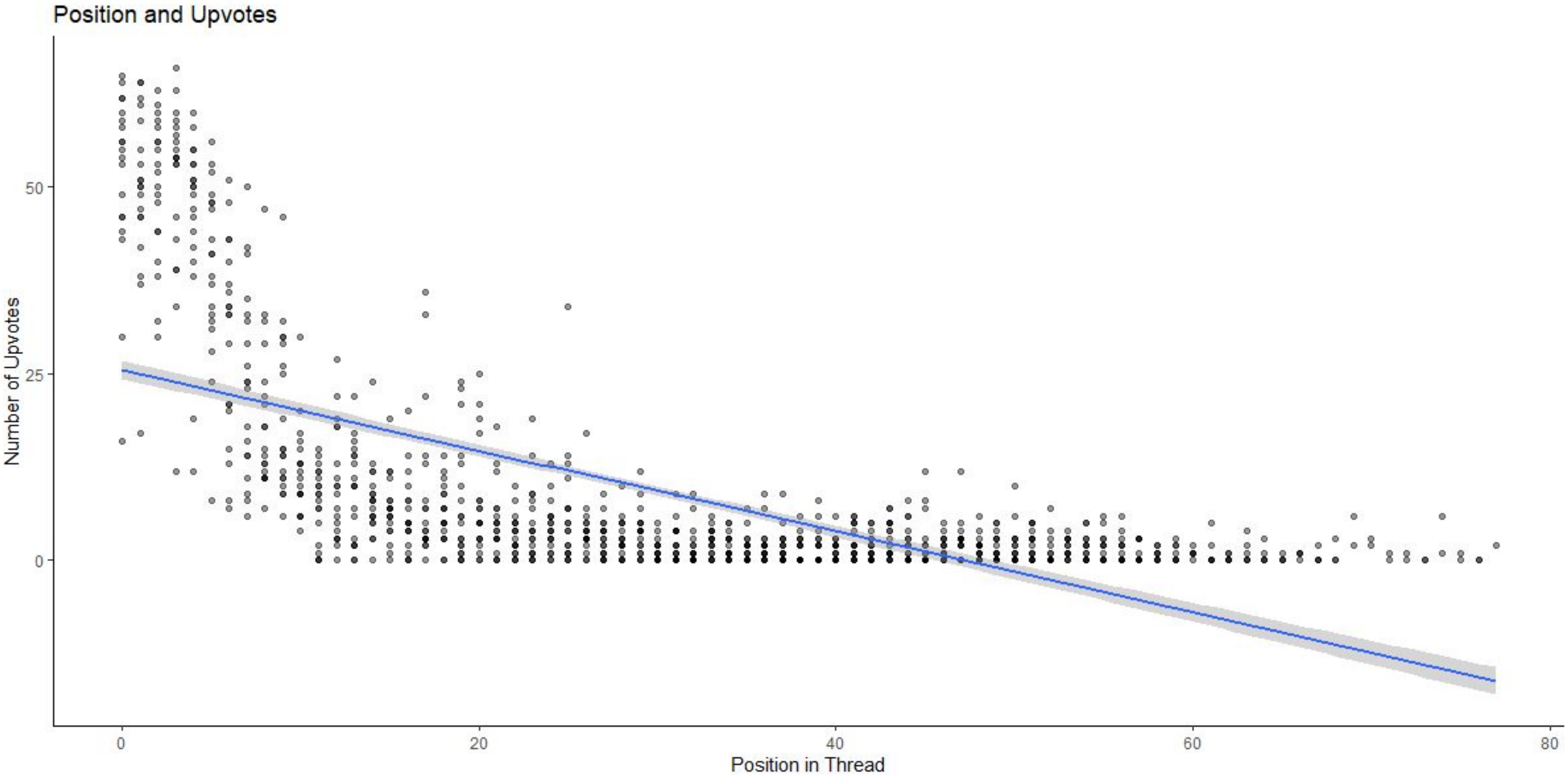
Conclusion 0

- The questions were distinct, unless there was a common theme between the talks!
- We read the materials and asked distinct questions (self-pat in the back)

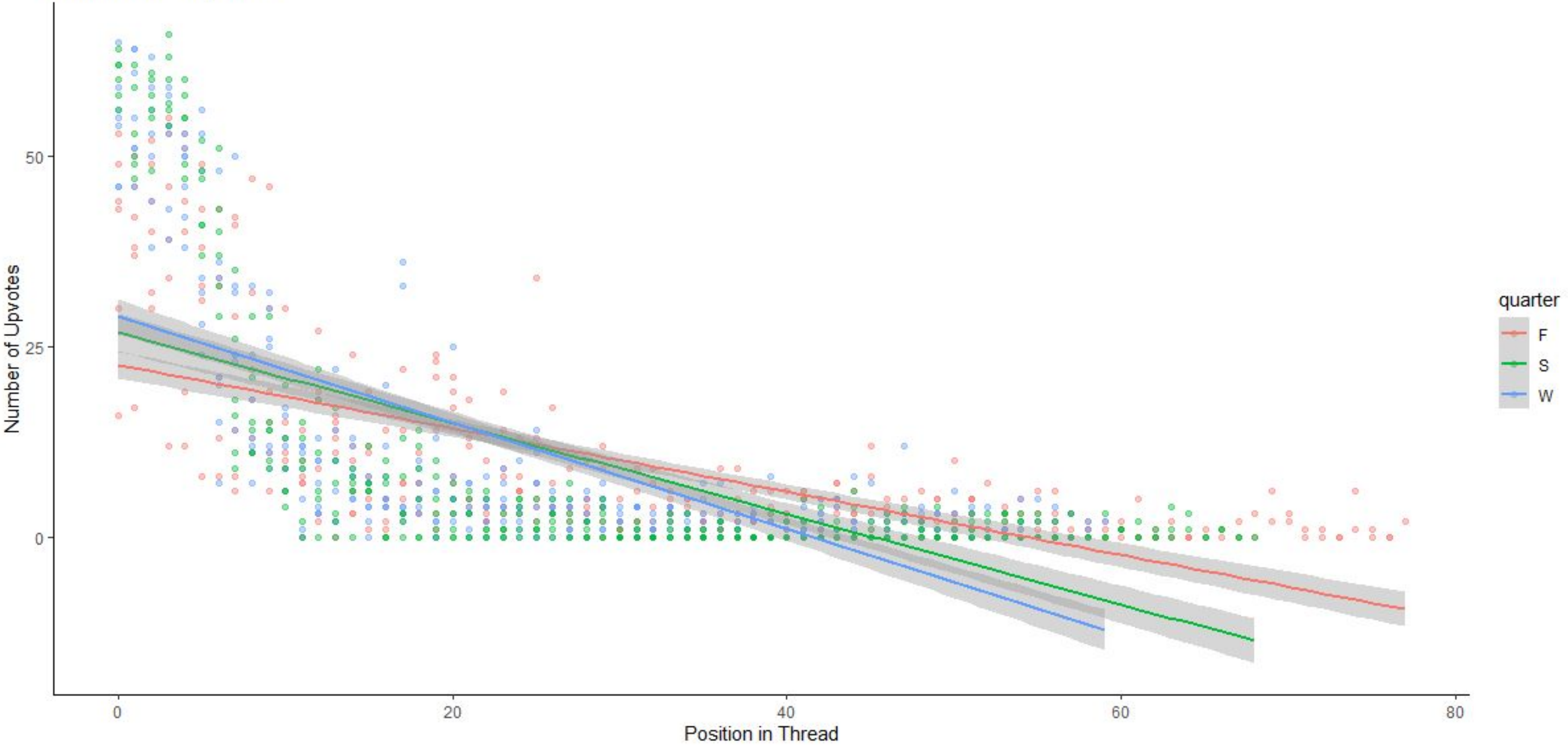
Statistical Analysis of the “Primacy Effect”

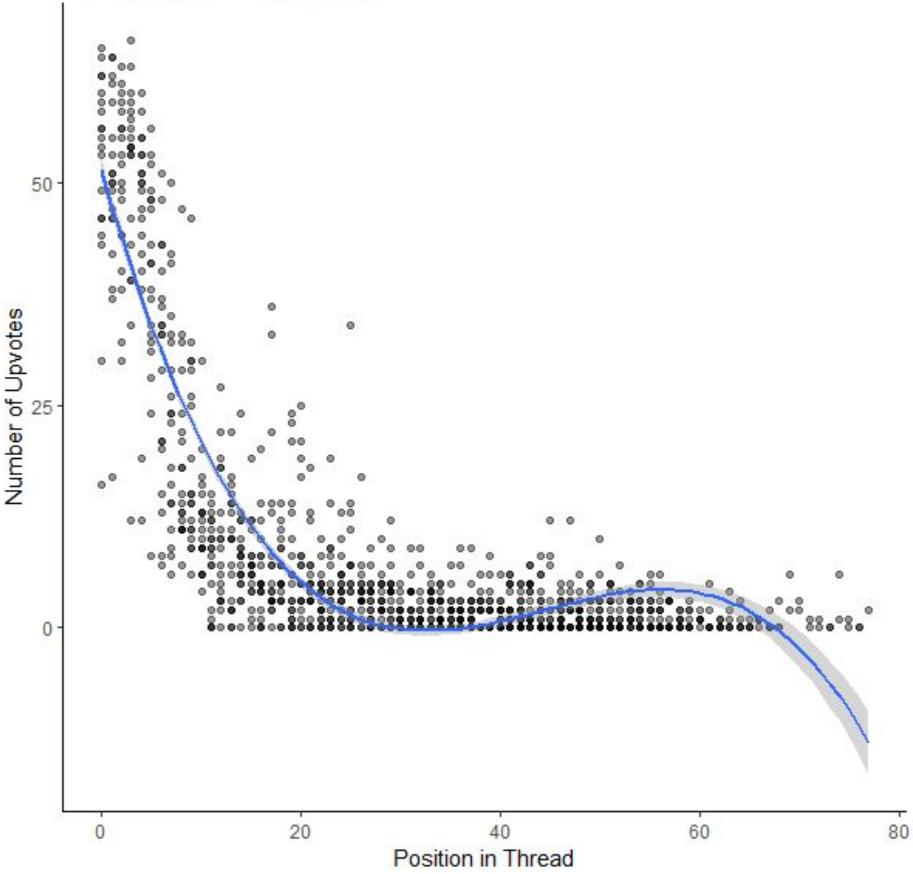
The “Primacy Effect”

- Bhargav (a MACSS alumnus who graduated this summer) suggested in the last year’s workshop mixer that people upvote questions that were asked first, not the best questions.
- I call this the “primacy effect”, a name I borrowed from memory science
- Makes sense, but we are quantitative people; **does the data show this?**

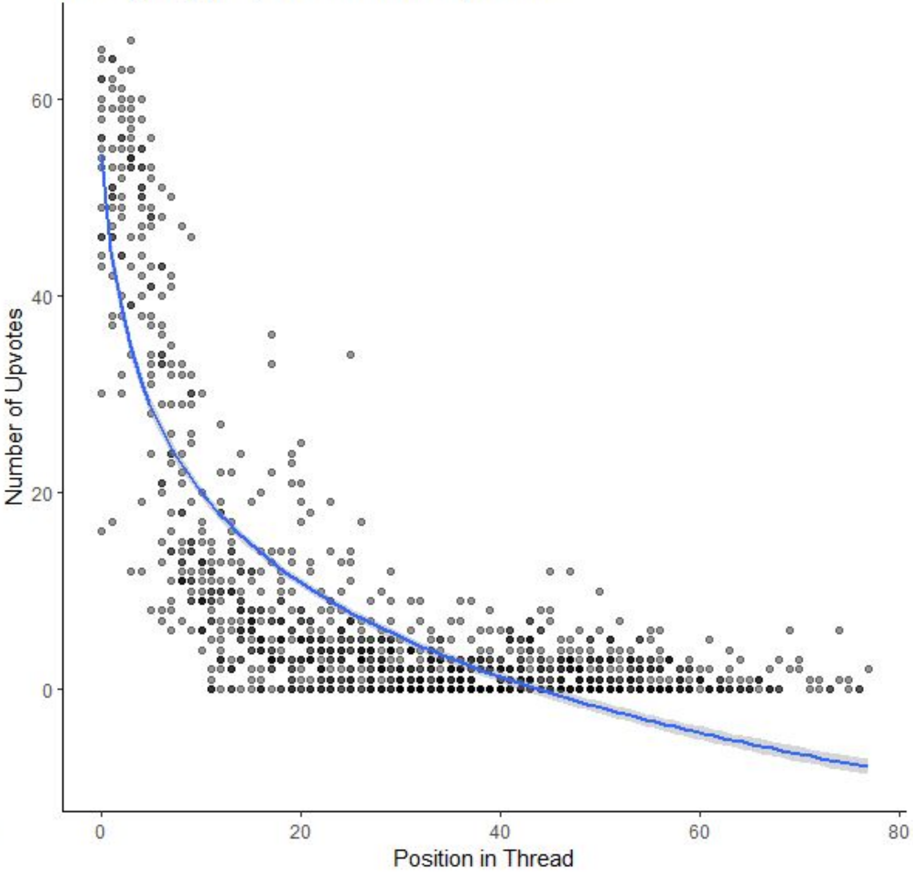

Position and Upvotes
Number of Upvotes
0
25
50
Position in Thread
0
20
40
60
80

Position and Upvotes



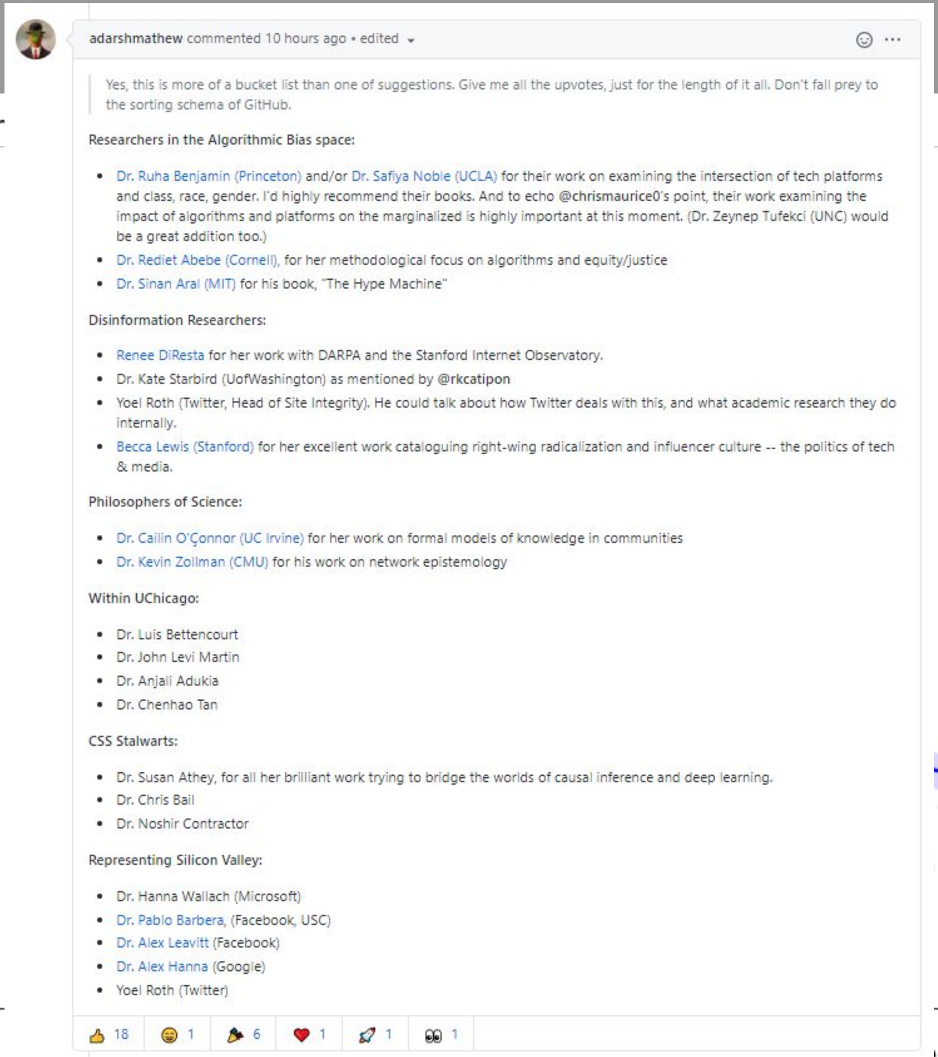


Third Degree Polynomial



Taking Log(x+1) on Number of Upvotes

Prir



> Yes, this is more of a bucket list than one of suggestions. Give me all the upvotes, just for the length of it all. Don't fall prey to the sorting schema of GitHub.

Researchers in the Algorithmic Bias space:

- Dr. Ruha Benjamin (Princeton) and/or Dr. Safiya Noble (UCLA) for their work on examining the intersection of tech platforms and class, race, gender. I'd highly recommend their books. And to echo @chrismaurice0's point, their work examining the impact of algorithms and platforms on the marginalized is highly important at this moment. (Dr. Zeynep Tufekci (UNC) would be a great addition too.)
- Dr. Rediet Abebe (Cornell), for her methodological focus on algorithms and equity/justice
- Dr. Sinan Aral (MIT) for his book, "The Hype Machine"

Disinformation Researchers:

- Renee DiResta for her work with DARPA and the Stanford Internet Observatory.
- Dr. Kate Starbird (UofWashington) as mentioned by @rkcatipon
- Yoel Roth (Twitter, Head of Site Integrity). He could talk about how Twitter deals with this, and what academic research they do internally.
- Becca Lewis (Stanford) for her excellent work cataloguing right-wing radicalization and influencer culture -- the politics of tech & media.

Philosophers of Science:

- Dr. Cailin O'Çonnor (UC Irvine) for her work on formal models of knowledge in communities
- Dr. Kevin Zollman (CMU) for his work on network epistemology

Within UChicago:

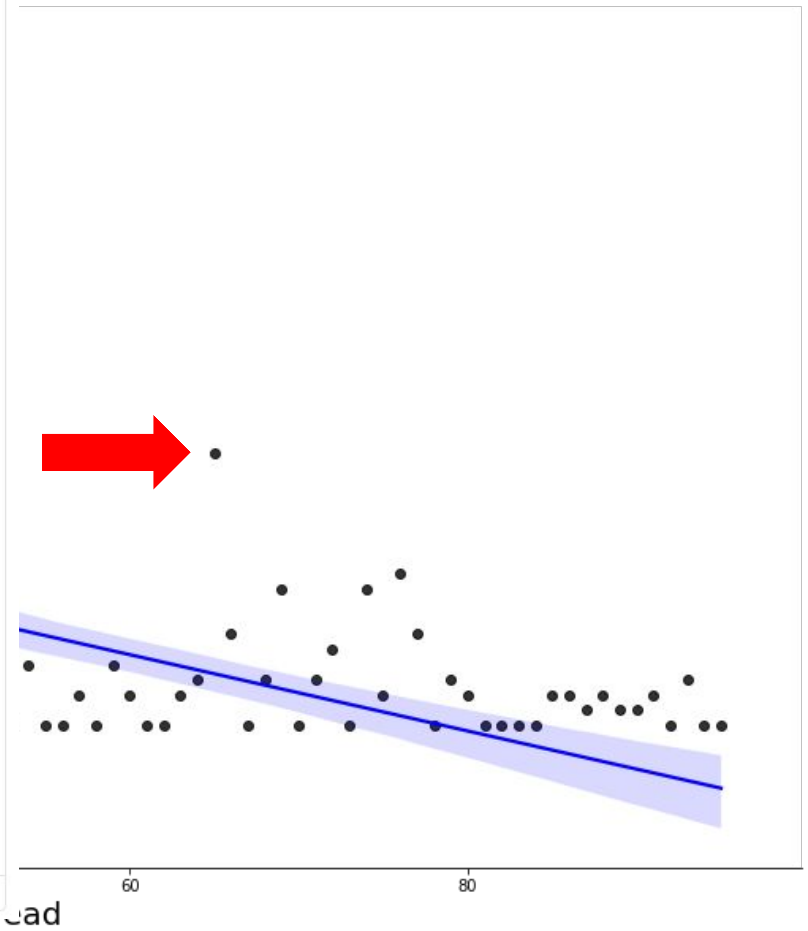
- Dr. Luis Bettencourt
- Dr. John Levi Martin
- Dr. Anjali Adukia
- Dr. Chenhao Tan

CSS Stalwarts:

- Dr. Susan Athey, for all her brilliant work trying to bridge the worlds of causal inference and deep learning.
- Dr. Chris Bail
- Dr. Noshir Contractor

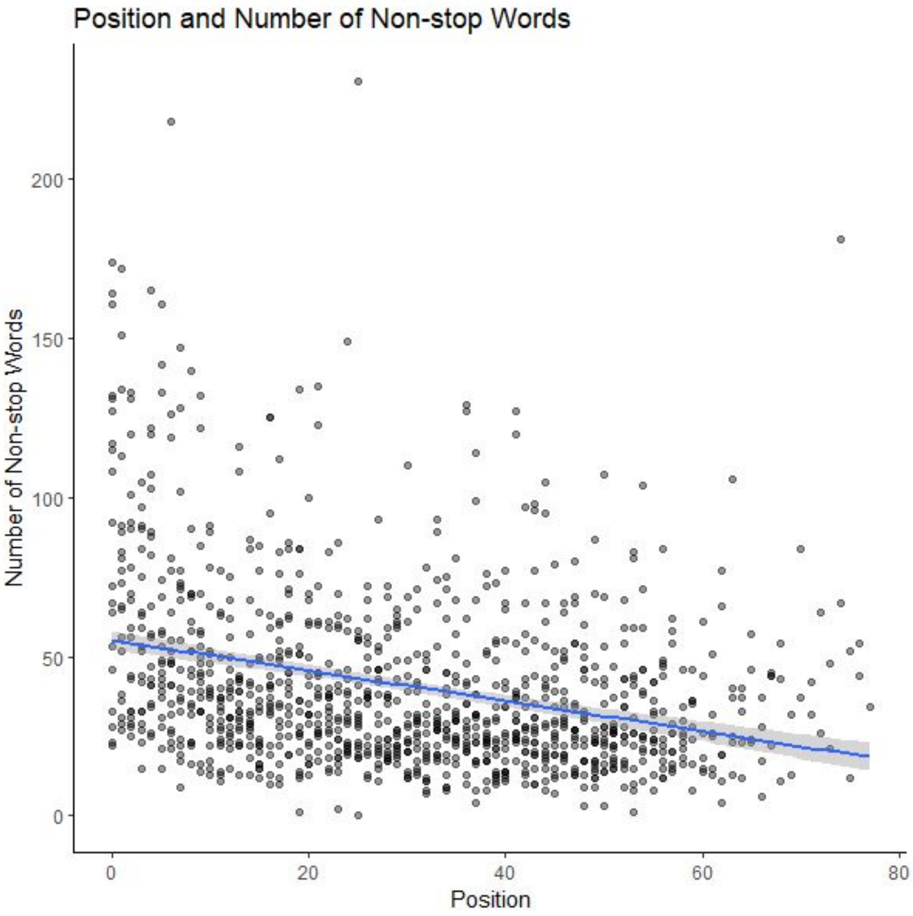
Representing Silicon Valley:

- Dr. Hanna Wallach (Microsoft)
- Dr. Pablo Barbera, (Facebook, USC)
- Dr. Alex Leavitt (Facebook)
- Dr. Alex Hanna (Google)
- Yoel Roth (Twitter)

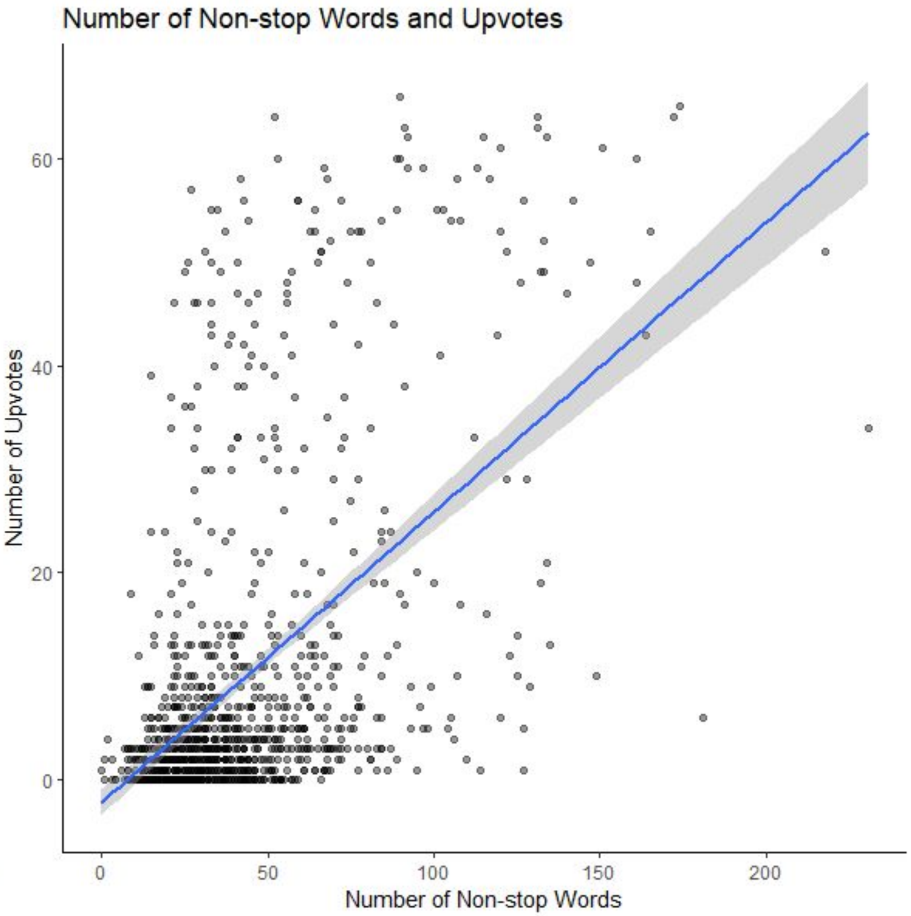


But maybe the quality of question is embedded in the order

- When I showed preliminary data on the winter quarter last year, Professor Evans suggested that good questions will be posted faster
- Simplest measure of quality (at least effort): number of non-stop words (e.g. “the”, “is”) in the question



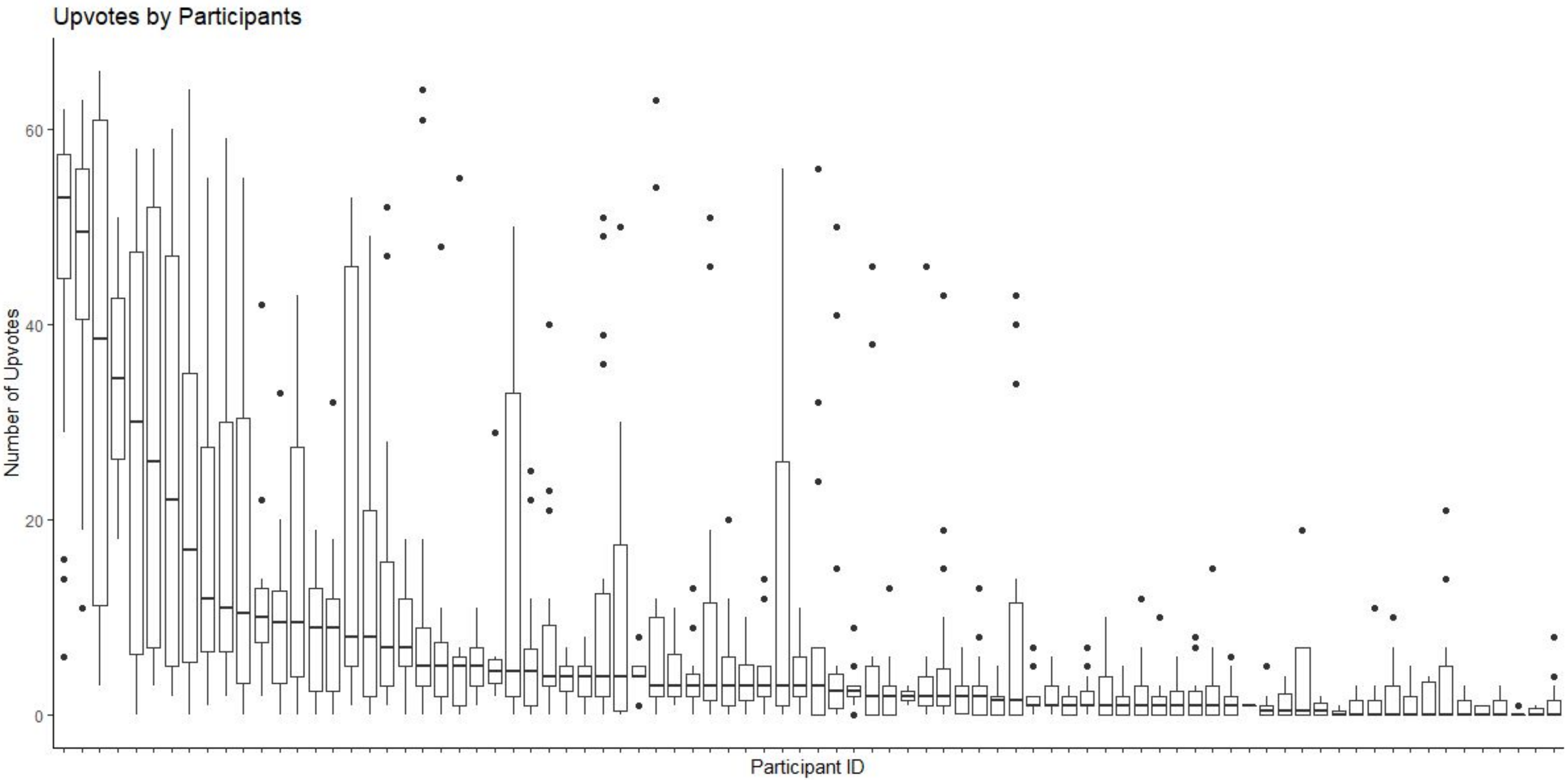
Position and Number of Non-stop Words



Number of Non-stop Words and Upvotes

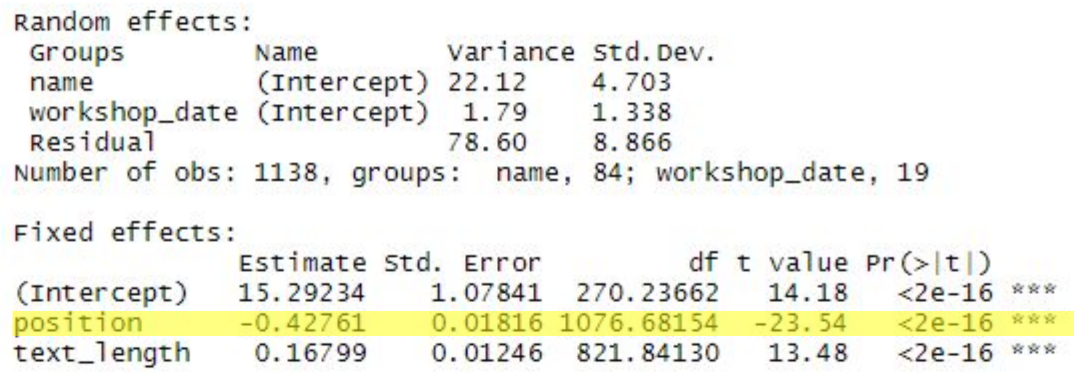
Or maybe it is a participant level effect

- Maybe some participants just get a lot of upvotes and that participants post faster…
- (of course this was not an controlled experiment, and I really do not want to get into the mess of causality)


Upvotes by Participants
Number of Upvotes
60
40
20
0
Participant ID

Does the effect survive in a complex model?

- Linear Mixed Model can isolate the effect of interest while controlling for the differences between participants (Baayen et al., 2008)
- Formula for the final model: `Number of Upvotes ~ Position + Number of Non-stop words + (1|Participant ID) + (1|Workshop ID)`



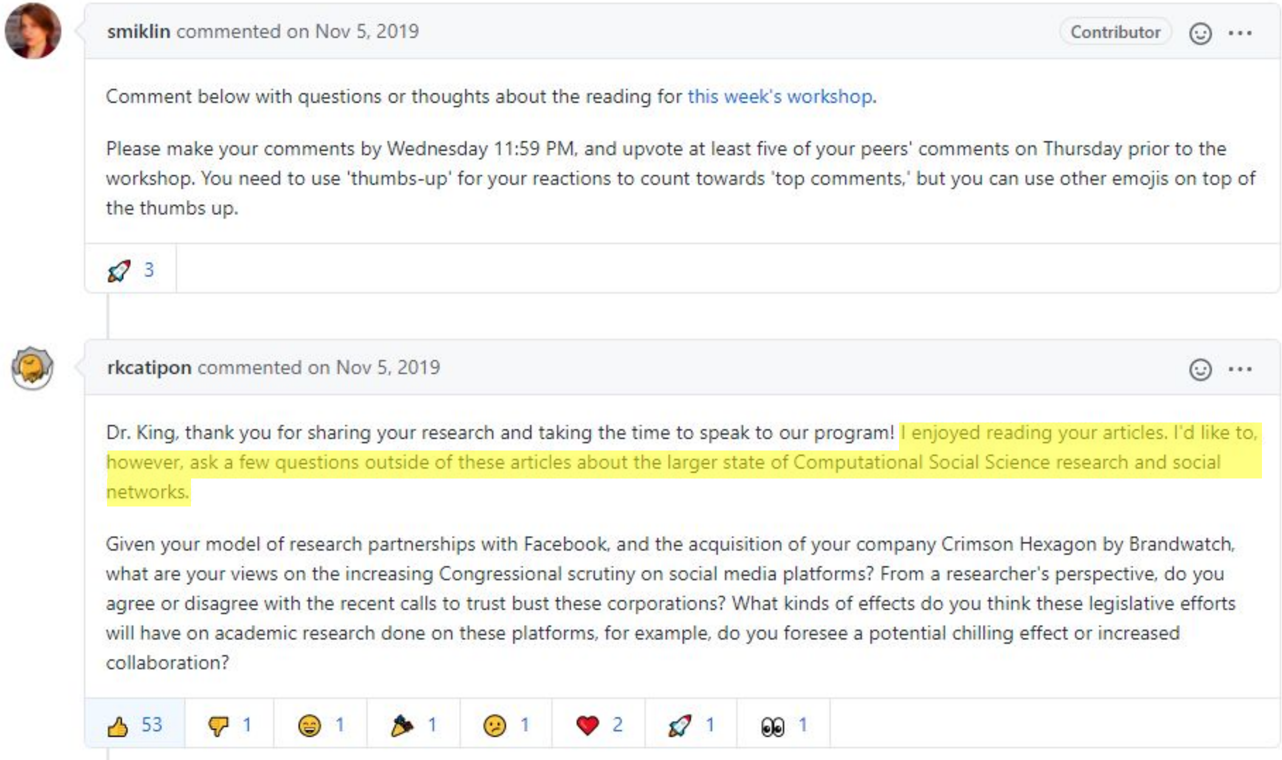
```
Random effects:
 Groups       Name        Variance Std.Dev.
 name         (Intercept) 22.12    4.703
 workshop_date (Intercept)  1.79   1.338
 Residual                 78.60    8.866
Number of obs: 1138, groups:  name, 84; workshop_date, 19

Fixed effects:
              Estimate Std. Error         df t value Pr(>|t|)
(Intercept)   15.29234    1.07841  270.23662   14.18   <2e-16 ***
position      -0.42761    0.01816 1076.68154  -23.54   <2e-16 ***
text_length    0.16799    0.01246  821.84130   13.48   <2e-16 ***
```

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, *59*(4), 390-412. https://doi.org/10.1016/j.jml.2007.12.005

Conclusion 1

- Primacy effect is a thing, even with several controls
- But number of non-stop words (very, very crude way of capturing “quality”) is a significant variable, too
- Personal thought: maybe even acknowledging this could change the pattern of workshops this academic year!

Game the system if you need to



smiklin commented on Nov 5, 2019

Comment below with questions or thoughts about the reading for this week's workshop.

Please make your comments by Wednesday 11:59 PM, and upvote at least five of your peers' comments on Thursday prior to the workshop. You need to use 'thumbs-up' for your reactions to count towards 'top comments,' but you can use other emojis on top of the thumbs up.

🚀 3

rkcatipon commented on Nov 5, 2019

Dr. King, thank you for sharing your research and taking the time to speak to our program! I enjoyed reading your articles. I'd like to, however, ask a few questions outside of these articles about the larger state of Computational Social Science research and social networks.

Given your model of research partnerships with Facebook, and the acquisition of your company Crimson Hexagon by Brandwatch, what are your views on the increasing Congressional scrutiny on social media platforms? From a researcher's perspective, do you agree or disagree with the recent calls to trust bust these corporations? What kinds of effects do you think these legislative efforts will have on academic research done on these platforms, for example, do you foresee a potential chilling effect or increased collaboration?

👍 53 👎 1 😄 1 🎉 1 😕 1 ❤️ 2 🚀 1 👀 1

Thank you for listening!

- Slides / code / data available on my GitHub repository: https://github.com/nwrim/CSSWorkshop_analysis
- Special thanks to Prof. Evans' course **Computational Content Analysis** (and wonderful TAs Bhargav and Hyunku) and Prof. Clindaniel's course **Computation and the Identification of Cultural Patterns** (and wonderful TA Abhishek)